# Todo

* Finish physics analysis section
* Pare down to essential questions to the extent possible
* Consider replacing multiple regressions w/ appropriate Bayesian approach to get more useful intervals (confidence intervals on coefficients not terrific)
* Consider a planned flowchart for analysis decisions - e.g., would we want to look at salience values in isolation first to finalize measures to use for preferential looking?
* Plan for deciding whether some events or event pairs are simply not worth including? (e.g. because we think maybe kids can’t see them well enough, etc.)
* **Plan for weighting preference data based on total looking time**
* Look at overall consistency (pref for particular stim) as an individually varying trait?
* Add: Use rate of switching (left->right)? (May not want to plan to do this if we need to do realtime rather than frame-by-frame coding for practical reasons.)
* Specify default significance level (& 1 vs. 2 tailed tests), type of sum of squares for ANOVA
* Treat object, flip, etc. as random effects?
* Literature review: findings of familiarity vs. novelty preferences in side-by-side designs?; known individual differences in looking patterns (see e.g. introduction to Barbaro, Clackson, & Wass 2016: <http://onlinelibrary.wiley.com/doi/10.1111/cdev.12689/full>). Findings of familiarity vs. novelty preferences on each event type? (If expecting genuine familiarity preferences, also need to adjust dependent measures/analysis to allow strength of preference, not just preference itself, to be predicted by attention…)
* Look at ManyBabies draft to check for things forgotten here...
* Clean up text & actually preregister. Could ideally submit as registered report, but for good reason unlikely to be accepted with ongoing data collection :)

Some pragmatic questions we might want to answer (not necessarily already in analysis plan yet):

* Do babies have consistent side biases? Distribution of directions? Distribution of magnitudes, allowing for switches between sessions? Does it matter which side they’re being held on?
  + One very practical exclusion question we might be able to address is “say we get a baby who spends >90% of time looking to one side across X trials or X minutes… how likely is it that that baby has a strong side bias and/or very “sticky” looking pattern, vs. this happening by chance or due to genuine preferences?”
* Approx. magnitude of noise due to individual, session, trial #, particular stimuli, mood/attentional state
* How does noise in preference depend on total looking time? How much less reliable are measurements with e.g. 5s total looking vs. 20?

# Preferential physics study overview

# 

The idea is to use dense sampling of individual infants on Lookit to conduct a detailed assessment of understanding of several physical principles. How stable are individual components across sessions and how independent? What does partial knowledge look like at the individual level?

We should be able to get reasonable bounds for each infant’s preferential looking ratio to simple violations of:

* Gravity: Completely unsupported objects should fall down immediately, rather than moving up, continuing in their current trajectory, or moving down at some delay.
* Inertia: Objects should continue roughly in their current trajectory when gravity is not a factor, rather than stopping and starting or turning around.
* Support: which of the following should fall (vs. stay put) after being placed?
  + An object placed mostly on the anchor
  + An object placed only slightly on the anchor
  + An object touching the side or bottom of the anchor
  + An object near the anchor but not touching it

Big picture: Why study individual behavior in more detail--what do we lose in studying groups of kids? We don’t know the extent to which a success means that a significant fraction of kids this age can do this task VS that all kids of this age can do this task to a significant degree. And even looking at the distribution of scores doesn’t clear this up, unless it’s an incredibly strong result (e.g. all kids get 9-10 of 10 questions right, or no kids get more than 6 out of 10). This matters for

* understanding how abilities are related to each other: we can’t get nearly as much out of age-based progressions without knowing how the noise works—especially for results of the form “n-month-olds, but not m-month-olds, can do X”
* understanding what partial knowledge & mechanisms of change in a domain look like: when kids “fail,” is that some kids making a correct prediction and some making an incorrect prediction? Or are they all failing to make any prediction and/or predicting at chance? When kids succeed but not at ceiling, are some getting one aspect and some another?

It may be that some kids don’t **express** nearly-universal knowledge on the dependent measures we collect. We can evaluate this explanation for noise and/or individual differences by studying the methods themselves, and kids’ behavior on them: e.g., can we predict the types of preferential looking responses we get from kids based on control tasks? How stable are those controls and task performance across sessions? (Especially interesting would be differences within kids in expression: kids may genuinely express knowledge at some times, but not others, due to attentional/emotional state changes.)

Especially in development, where we’re interested in the underpinnings of human cognition, the difference between “some babies use this type of information but others do something else” and “all babies have this type of information available to them, unless something’s wrong” matters a lot—this is exactly where we care about universality. But of course infant methods are hard to collect repeatedly compared to, say, educational data later on--which is where Lookit has an advantage.

# Experiment protocol

Children complete 24 20-second preferential looking trials per session; families are encouraged to complete 15 sessions within 2 months. Parents complete a short mood survey and go through some instructions before the preferential looking portion. They’re asked to hold their children looking over their shoulders during this portion to avoid parental bias. Parents can end the study at any point and skip to the post-study survey.

Each trial begins with an object intro ([video of Kim saying “Look, this is a …”](https://s3.amazonaws.com/lookitcontents/exp-physics-final/examples/0_introsA.mp4) and demonstrating use of an object – e.g. biting into an apple, putting on hand lotion, drawing with a marker, eating with a spoon) that lasts about 5 seconds. This is intended as an attention-getter to re-orient children towards the center, while in principle reinforcing that the object in question is not an agent and should be expected to follow normal physical laws. Then two events involving that object are shown simultaneously, one on the left and one on the right, looping for 24s (event videos range from about 2-5s). Events always show the same object, same camera angle, same background, with a difference only in the “outcome.” Event types are shown in the table below; each concept is presented 4-6 times, with 2-3 repetitions of each event type.

Although events are short, they loop continuously for 20s. Realtime events are shown so that “expected” events are at natural speeds, and not potentially seen as violating physical principles due to happening too slowly. Although this means that objects fall quickly, children learn from realtime experience in everyday life.

[Note: links are to “collages” of stimuli; children are always shown just two event videos at a time. If you can’t view the .mp4 videos in your browser, just replace the extension with .webm.]

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Concept** | **Event** | **Outcome** | **Objects** | **Camera angles/BGs** | **Times shown** |
| **Gravity** | [**Table**](https://s3.amazonaws.com/lookitcontents/exp-physics-final/examples/3_table.mp4): Object is rolled/slid off a table and continues… [[example pairing]](https://s3.amazonaws.com/lookitcontents/exp-physics-final/stimuli/table/mp4/sbs_table_continue_down_lotion_c2_1_NR.mp4) | **down** | Apple, cup, lotion, orangeball, whiteball | 2/1 | 2 |
| **up** |
| **horizontal** |
| [**Toss**](https://s3.amazonaws.com/lookitcontents/exp-physics-final/examples/2_toss.mp4): Object in hand is… | Tossed up, falls **down** | Apple, cup, lotion, orangeball, spraybottle, whiteball | 2/1 | 2 |
| Tossed down, falls **up** |
| [**Ramp**](https://s3.amazonaws.com/lookitcontents/exp-physics-final/examples/1_ramp.mp4): Object is placed in center of ramp and is released to roll/slide… | **down** | Apple, cup, lotion, spraybottle, whiteball | 2/2 | 2 |
| **up** |
| **Inertia** | [**Stop**](https://s3.amazonaws.com/lookitcontents/exp-physics-final/examples/4_stop.mp4): Object rolls from one side of screen and stops in the middle before continuing | stopped/restarted by a **hand** | Block, flashlight, marker, sunglasses, toycar, train | 2/1 | 2 |
| stops/restarts on its own, with **no hand** |
| [**Reverse**](https://s3.amazonaws.com/lookitcontents/exp-physics-final/examples/4_reverse.mp4): Object rolls/slides from one side of the screen and collides with… | **barrier** | 2/1 | 2 |
| **no barrier** (object follows same trajectory as above, but doesn’t actually touch the barrier) |
| **Support** | [**Fall**](https://s3.amazonaws.com/lookitcontents/exp-physics-final/examples/5_fall.mp4): An object is placed […] a cabinet and immediately falls | **mostly on** | Book, brush, duck, hammer, shoe, tissues | 1/1 | 3 |
| **slightly on** |
| **next to** |
| **near** |
| [**Stay**](https://s3.amazonaws.com/lookitcontents/exp-physics-final/examples/6_stay.mp4): An object is placed […] a cabinet and stays there | **mostly on** | 1/1 | 3 |
| **slightly on** |
| **next to** |
| **near** |
| **Control** | [**Same**:](https://s3.amazonaws.com/lookitcontents/exp-physics-final/examples/7_control_same.mp4) Distinguishable but similar physically-possible human actions on objects, like rotating an object about one axis vs. another | version **A** | Box, eraser, funnel, scissors, spoon, wrench | 1/1 | 3 |
| version **B** |
| [**Salience**](https://s3.amazonaws.com/lookitcontents/exp-physics-final/examples/7_control_salience.mp4): Physically-possible human actions on objects, some more interesting, like flipping a spoon vs. slowly extending it or erasing a drawing vs. an empty board | **boring** | 1/1 | 3 |
| **interesting** |
| [**Calibration**](https://s3.amazonaws.com/lookitcontents/exp-physics-final/stimuli/attention/mp4/calibration_LR.mp4): spinning ball moves back and forth | [Entire screen, start left/right] | N/a | n/a | 2 |

Parents can pause individual trials. If they pause during the intro, they just start over upon restarting. If they pause during the test (up to one time per trial) they restart from the intro, but then the left and right test videos are switched for the test phase.

*Trial order*: Trials cycle through gravity, inertia+calibration, support, and control (same/salience) pairings during a session; the order of these concepts is chosen from a list and changed (cycling through a list of orders) each session. There are six videos shown in each category in total.

Within each category, objects are assigned to comparison types (e.g. “apple” assigned to “table, down vs. up”) by choosing from a list of acceptable mappings, again incremented per session. (The first session value was selected randomly from the first six options initially [by accident], and is now selected from all possible mappings.)

There are six possible comparisons for the stay and fall events; three comparisons are assigned to stay and three to fall, with the selection again cycling through a random list of such assignments per session. Left/right placement, horizontal flipping of the left and right events, camera angles, and backgrounds are chosen randomly with the constraint that half of the ‘more probable’ events are on the left within each category. Calibration trials (grouped with inertia videos for purposes of assigning object intros) are placed at trials 3 and 6, so that they are always available for kids who completed enough trials for the session to be included (and so that if there are differences in coding quality across trials, we’re not excluding on the basis of when calibration happened).

# 

# 

# Overview of data collected

Data for each child:

* Age at start
* Demographic info (optional, typically reported): family income, languages spoken at home, parent education level(s), number of parents in the home, number of children’s books in the home, child’s race, number and age of siblings, country + US state if in US, urban/suburban/rural

Data for each session:

* Child’s age
* Number of previous sessions completed
* Time since last session
* Mood data (Before beginning study, by parent report. Scales 1-7: CHILD: tired-rested, sick-healthy, fussy-happy, calm-active; PARENT: tired-energetic, overwhelmed-ontopofthings, upset-happy. How long since child woke up, how long since child ate, how long until child is due for nap/sleep; what child was doing before this.)

Data for each trial:

* Time looking L, R (time sequence of looks left, right, away)
* # Fixations
* Parent behavior: times of talking, pointing, and peeking
* Infant behavior: times of fussing & rating of fussiness level low or high
* Trial #

Predictors for fraction time to left

* Comparison type (e.g. ramp up vs. down), nested within
* Event type (e.g. ramp), nested within...
* Concept (gravity, inertia, support)
* Object
* Whether each side is unexpected (i.e., does the event on the left clearly violate a physical principle? Does the event on the right? Sometimes both are unexpected to adults, e.g. when an object near a cabinet and an object next-to a cabinet stay put; sometimes neither is. In some cases this depends on the child’s potential beliefs, see modeling…) & which side is \*less\* expected

# 

# Video coding

Completed coding is to nearest frame in VCode (coder manual available), with some pilot data coded both closer to realtime (not adjusting looks to the nearest frame) and to nearest frame by the same coder to check reliability. Coding a complete session takes around an hour for an experienced coder to do frame-by-frame.

# Recruitment: age range and sample size

Age range 4-12 months at start of study; continue for up to 61 days; target ‘complete’ dataset is 12 usable sessions. (No major age differences in data quality or salience/same controls seen in piloting.)

Plan to recruit as large a balanced sample as practical given time constraints on both testing and recruiting. Because parent interest is highly unpredictable for a new online study, stopping criterion may be reaching 50 participants with at least 12 usable sessions each OR a specified target end date. Could also aim for X with at least Y usable sessions each AND at least Z usable sessions total (to allow for various possible distributions of sessions). All recruitment decisions are / have been made without examining dependent variables.

All partial datasets (<12 sessions) and any extra data collected (due to inclusion depending on coding) will be included in the analysis.

# 

# Analysis plan

## Exclusion criteria:

Child

* Gestational age at birth < 37 weeks, for any analyses using age. Unknown gestational age will be used but prevalence reported. (Followup to check that inclusion of premature infants does not qualitatively affect other results, and/or to display results from premature infants with adjusted/non-adjusted age - exploratory.)
* Children who participated in the pilot study
* Children whose parents spontaneously report developmental/medical issues that would likely explain some differences in task: vision or hearing impairment; cognitive or neurological disorders including due to trisomies.
* Include data from children with any number of sessions (will probably have many with <15 in addition to “complete” data); analyses described should be able to handle this appropriately.

Session

* Sessions where children are outside age range of 4-14 months, except for binned-by-age analyses where we may display data from children outside the age range (without affecting any other values) if we end up having it. Adjusted age will be used for premature infants.
* Sessions with < 6 trials (& don’t count as a session for session # purposes). Parents are encouraged not to complete sessions within 6 hours of each other. For each session (process later -> earlier), if another session with fewer completed trials happened within six hours, use this one instead.This leads to reasonable outcomes even in the unlikely event someone’s doing the study every 5 hours.
* Require calibration performance >75% to use session. Calibration scores seem to be mostly due to difficulty coding and might therefore index overall confidence in other judgments; timing differences in webcam stream vs. displayed stimuli will also affect calibration. Pool all looking across the two calibration trials to compute an overall calibration score, so that if kids aren’t looking as much for one of the trials we don’t average in a much noisier measurement. Score is fractional looking time to correct side during the middle of periods when the ball should be static: [0.5, 3.5], [5.5, 8.5], [10.5, 13.5], [15.5, 18.5], [20.5, 23.5].
* If the participant has <12 usable sessions spread over >60 days, use sessions from the earliest 60-day period (inclusive) with the most usable sessions. If the participant has >=12 usable sessions over >60 days, use sessions from the earliest 60-day period (inclusive) with at least 12 sessions.
* Where absolute session number (as an index of how many times the child has seen stimuli, etc.) is relevant, assign the first session used in analysis a session number according to the number of ‘experienced sessions’ in the preceding 60 days. Experienced sessions include sessions where the child is out of age range or calibration performance is poor, but not sessions with <6 trials. (For instance, if a child participates at 4 months of age and then 12 times from 10-11 months, the latter set of data is used and the first session at 10 months is considered session 1. If a child participates on days 1, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, then sessions starting at day 30 are used but the first session number is 2.

Trial

* Require >= 2s looking to use a trial. Don’t otherwise deal with shorter/longer LTs except in (exploratory) model.
* Omit periods where the child is out of frame or gaze is otherwise impossible to code. Treat as out of frame any periods where the video is ‘frozen’ for >1s (start treating as out of frame 1s after this period begins) and report prevalence.
* Exclude trials where child is fussing >50% of the trial. In contrast to omitting periods where child is fussing, this avoids dependence on exactly which frames are considered fussy, at the cost of a threshold effect we expect not to affect many trials. >50% applies to length of video, not otherwise codable data (i.e. we allow fuss coding during outofframe periods).
* Parent interference: Exclude *periods* of trials after parent peeks, points while peeking, or speaks in any way that could bias child (“what’s that ball doing?” but not “keep looking sweetie!”) Include periods where the parent’s eyes are not visible and we can’t tell where they’re looking (unless there is reason to believe they are looking) and periods where the parent is looking away but may see stimuli peripherally.

## Measures

* Fractional looking time (fLT) measurement (looking time to target / (target + distractor)), starting 4s into trial to wait for a stable preference (conservative, based on pilot data). Some periods of trial may be omitted as described in exclusion criteria; after omitting these periods, 2s minimum looking time criterion per trial is applied. [Note that current processing of videos extracts the last 20s of a 24s trial for coding.]
* |fLT - 0.5|, non-directional measure of strength of preference, for ‘same’ events only
* Rate of switches throughout *session* (# switches R-L or L-R / total looking time).
* Do expect some dependence on trial #. Use all trials; assume that changes over time are shared across kids, and include trial number in regressions where appropriate.

## Stability, variation, & controls

* **General markers of looking patterns + choice of measures.** 
  1. Measures (use at individual trial level):
     1. Side bias: fractional looking time to R during ‘same’ events (3 events/session)
     2. Stickiness: |fLT – 0.5| during ‘same’ events (3 events/session)
     3. Sensitivity: fLT to more interesting during salience (3 events/session)
     4. Total looking time (all events)
  2. How stable are control measures across sessions, and do they change with age?
     1. Partition variance for each measure: measure ~ age + (1|child/session). (Fraction of **total** variance explained = intraclass correlation coefficient.) Report coefficient of age & test for significance.
     2. Display overall distributions of these measures (one mean per child) and plot measures against sessions (e.g., one line per child)
  3. Are controls well predicted by mood measures? Regress each control measure (using means per session, and |(mean fLT to right) - 0.5| for side bias as a measure of strength of bias to either side) using model: measure ~ 1 + parentscore + childscore + childactivity + timesincewaking + timeuntilsleep + trial# + (1|child). Report overall significance of model (based on F value) and coefficients of individual predictors. Predictors:
     1. Parent score: mean of z-scored parent items
     2. Child score: mean of z-scored child items rested, healthy, happy
     3. Child activity: calm-active score
     4. Time since waking up
     5. Time until due for sleep (“overdue” = 0, “no schedule” = missing data.)
  4. How well does side bias during control stimuli predict side bias during tests? sb\_test ~ sb\_control + (1|child). Use mean looking time to right across all same/test trials in a session (sb\_control includes up to 3 ‘same’ trials, sb\_test includes all other trials except calibration). Report coefficient & test for significance of sb\_control.
  5. How well does sensitivity control predict fLT on events? fLT ~ sensitivity + (1|child). One mean fLT value per session, including all tests with unambiguous expected outcome to adults (excluding calibration, same, table up-continue, stay/fall without mostly-on as one of the outcomes). Report coefficient & test for significance of sensitivity.
  6. How well does stickiness predict noisiness of data? variance\_salience ~ stickiness + (1|child). Report coefficient & test for significance of stickiness.
* **How stable are kids’ looking patterns on test stimuli across sessions?** How much do the specific videos (e.g. object choices, backgrounds) matter in comparison to the event types?

fLT\_trial ~ concept/event/comparison + concept:event:comparison:object + concept:event:comparison:object:cameraangle + concept:event:comparison:object:background + concept:event:comparison:object:flip + trial# + (1|child/comparison) + (1|child:session). Report fraction variance explained by child/comparison & by session, coefficient for trial# with significance, and use ANOVA to report overall effects of object, camera angle, background, & flip across children.

* **Do preferences change over the course of the 15 sessions, across children?** fLT\_trial ~ concept/event/comparison + trial# + session# + (1|child/comparison). Report coefficient for session# and test for significance.

## Physics (in progress)

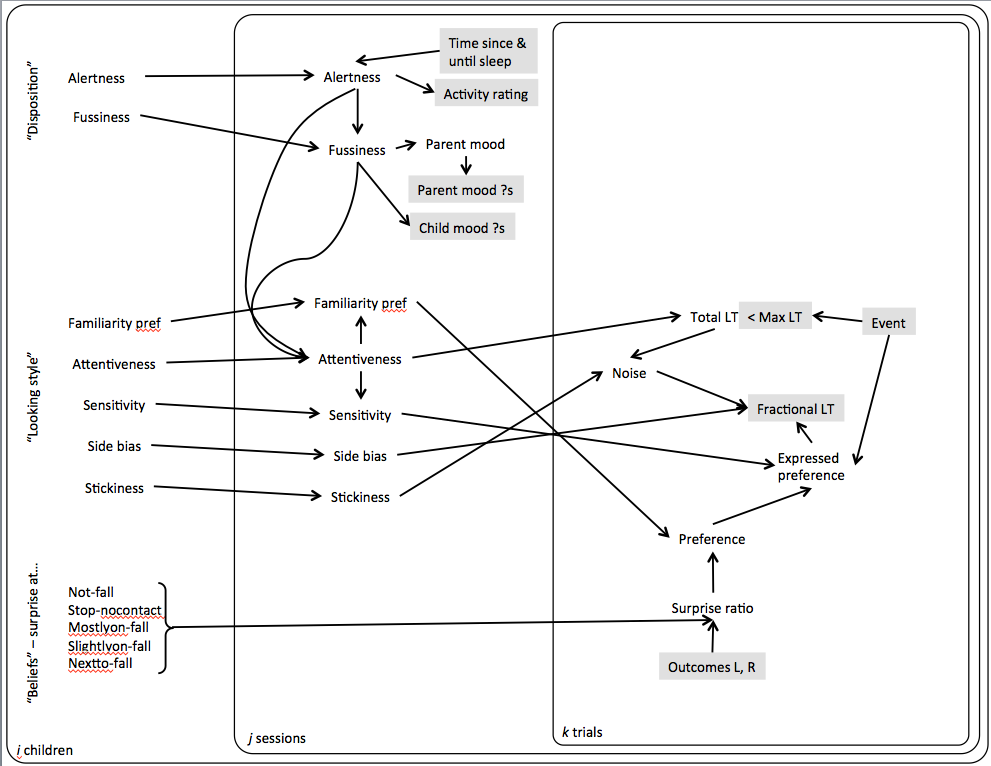
* Summary of children’s preferences on these events:
  1. Bin together each comparison type (each physics comparison + salience control), all measurements per child. Separately bin by concept (gravity, support, inertia). Add total time looking at target and distractor across all trials to get an fLT measurement. Bootstrap confidence intervals per child by resampling sessions, then trials.
     1. What fraction of children have a preference for either the expected or unexpected event with 95% CI not overlapping 50% looking, for each grouping? (Just informative for design of future studies - how much data do you need to see individual preferences? May also want to calculate for smaller subsets of the data, e.g. how many sessions until X% children show individually-significant results when group-level effect size is Y.)
     2. Regress mean fLT per comparison type on age (mean age across usable trials for this comparison type, per child). (Could also do mean fLT ~ age + comparisonType + (1|child) or similar, except for concerns about differing correlations among subsets of comparison types.) For each comparison type, check for significant effect of age and significant difference from chance at reference point of 12mo. Display age trends.
     3. Gravity, graded judgment: predict table up v down > down v continue > up v continue. Is each pairwise comparison significant?
     4. Support: How well do “stay” preferences anticorrelate with “fall” preferences? (6 comparison types per child). (exact test/estimation description here)
     5. Put together support measurements into stay minus fall preference on theoretical grounds (i.e. pref for slightly-on stay vs. mostly-on stay should be opposite & closely related to pref. for slightly-on fall vs. mostly-on fall). Intuitive approximate magnitude ordering based on adult understanding of physics: mostly-near, mostly-next, mostly-slightly, slightly-near, slightly-next, next-near. What orders do we actually see w/i kids?
        1. ~~Stage theory: Might make crude predictions based on which should stay vs. fall—if only mostly-on should stay put, then mostly-slightly, mostly-next, mostly-near are bigger differences and might show bigger preferences than the others, since in these cases we have one “expected” and one “unexpected” outcome. If mostly-on and slightly-on both stay put, then expect instead mostly-next, mostly-near, slightly-next, slightly-near to show bigger preferences than the others. Etc. Project preference vectors onto [1,1,1,0,0,0], [0,1,1,0,1,1], [0,0,0,1,1,1] and plot transformed coords by age, where transformed coords basically represent “how much like a partial-support-knower do you act,” “how much like an any-support-from-below-knower do you act,” “how much like an any-contact-is-support-knower do you act.”~~
        2. Overall, across kids, what preference vectors do we see, binned by age group? (plot mean vectors over time?) Could imagine getting closer to [1,1,1,0,0,0] as above or getting better on everything, even the ones that aren’t disparate on possibility (i.e. might expect a baby who “really gets it” to not care about something staying in midair vs. staying when placed next to a cabinet, because they’re both obviously impossible; OR might expect a baby who “really gets it” to differentiate even this subtle difference in probability because hey, maybe it’s a sticky table or my perception is noisy or something. Nothing in Baillargeon stage theory predicts getting better at this comparison with age, so it’s a nice test.
* Overall structure of tasks: Bottom-up “sort” by pairwise correlations (as with similarity)? -> correlation clustering on individual comparison types or dendrogram. (Need exact description here)

## Exploratory questions (data that may be useful to other researchers):

* **How do preferences change throughout individual trials?**
  + Plot: Preference for unexpected over course of trial, per comparison.
  + Plot: Preference for eventual preferred direction over course of trial, all physics events.
  + Plot: Rate of switching over course of trial, binned per second.
* **How often do children show familiarity rather than novelty effects?**
  + Can’t perfectly address this question, but related results may be useful for future experimental design. Note that any investigation of the predictiveness of the strength of a familiarity preference suffers from the confound that the lower the measurement, the higher the likelihood the child does NOT have a strong true novelty preference, even if there are only novelty preferences--so degree of “familiarity preference” is informative in any case.
  + DV: mean fLT per comparison type per child per session. LME: fixed effect of comparison, random effect 1|child/event. Display distribution of event (fixed) + child + child/event random effects, per event. (Only convincing pattern would include clear “bumps” both above and below a central cluster, which seems highly unlikely. Otherwise very hard to tell anything.)
* Check impact of noise due to varying calibration, parent interference, fussiness: Followup “conservative” analysis, see Frank Amso Johnson 2014).
* Use control measures to weight other measurements.

## Bayesian model to infer gravity, inertia, support “belief states” within children (exploratory, not part of this project)

Expect at least ~14K trials total, ~290/child, for very rough scale of # parameters we can infer. Due to large uncertainty in the contributions of and forms of relationships among various contributors to infant preferential looking, the model will be most productively treated as exploratory rather than as a primary way to infer estimates of latent variables & individual variation. A good outcome would be to be able to say, after describing what we see in the data, that e.g. “we’re *able to* model these responses by inferring these looking-style parameters, which vary primarily across sessions with relatively small individual differences.” In this view the model is primarily a proof-of-concept and starting point for future research.



### Latent variables - “looking style” (overall & per session):

* Sensitivity: extent to which differences in surprisingness of outcomes affects looking preference
* Side bias: tendency to look right
* Stickiness: tendency to strongly prefer one side
* Attentiveness: tendency to pay attention / processing speed; affects total LT, sensitivity, familiary preference
* Familiarity preference: degree to which predicted/familiar outcomes are preferred to unpredicted overall

### Surprise ratio & beliefs (this is one starting version, staying close to stimuli rather than broadly conceptual):

* Beliefs are continuous, representing how surprising a particular outcome is relative to another.
  + Not-fall: Unsupported object fails to fall, vs. falling.
  + Stop-nocontact: Object stops without contact force, vs. with contact
  + Mostlyon-fall: Object placed mostly on a surface falls, vs. unsupported object
  + Slightlyon-fall: Object placed slightly on a surface falls, vs. unsupported object
  + Nextto-fall: Object placed next to surface falls, vs. unsupported object
* Salience is inferred across children
* Force fall/stay surprise to be opposite
* Surprise ratio computed by dividing R by L outcome surprise values on chart below:

|  |  |  |
| --- | --- | --- |
| Event | Outcome | Surprise |
| Table | Up | not-fall |
|  | Down | 1 |
|  | Continue | not-fall |
| Toss | Up | not-fall |
|  | Down | 1 |
| Ramp | Up | not-fall |
|  | Down | 1 |
| Stop | Nohand | stop-nocontact |
|  | Hand | 1 |
| Reverse | Nobarrier | stop-nocontact |
|  | Barrier | 1 |
| Stay | Mostly-on | 1/mostlyon-fall |
|  | Slightly-on | 1/slightlyon-fall |
|  | Next-to | 1/nextto-fall |
|  | Near | 1 |
| Fall | Mostly-on | mostlyon-fall |
|  | Slightly-on | slightlyon-fall |
|  | Next-to | nextto-fall |
|  | Near | 1 |
| Salience | Interesting | salience |
|  | Boring | 1 |
| Same | A | 1 |
|  | B | 1 |

### Predicting fractional looking time:

Preference: Combines surprise ratio R, familiarity preference F.

Example: Preference = ([R / (1 + R)] - 0.5) \* F

F in [-1, 1] where 0 is indifference; R in (0, inf) where 1 is indifference. Preference in [-0.5, 0.5], -0.5 = preference for familiar, 0.5 = preference for unfamiliar.

Expressed preference: Combines preference, sensitivity, event. Values per event (table, toss, etc.) are fitted across kids to allow some events to be clearer/more comprehensible examples of underlying concepts.

Example: Expressed preference = preference \* sensitivity \* event

Fractional looking time: Combines expressed preference, side bias, & noise. (Lower-LT trials and higher-stickiness kids are noisier.)

### To do...

To add / consider adding:

* Age! (better to do only across kids?)
  + Effects on looking style
    - Age affects each parameter at session level, in combination w/ child’s baseline
  + Effects on beliefs
    - Age affects each belief at session level, in combination w/ child’s baseline
* Calibration (as another determiner of noise)
* Fatigue over trials
* Objects, camera angles, backgrounds
* # switches measure
* SES effects
* Planned comparisons of several models for how beliefs influence LT

To do:

* Priors
* Forms of relationships (including for per-kid parameters currently not specified as coming from a particular distribution)
* Plan for how we’ll use the inferred parameters...

# 

# 

# 

# Some messy power analysis/estimation notes:

Estimating variance of measurements of single task:

**Novelverbs** (2-year-olds, asked to find verbs):

Mean standard deviation within first verb: 0.14

Size of difference between consecutive measurements:

mean 0.17, IQR [0.05, 0.24]

Mean standard deviation within second verb: 0.20

Size of difference between consecutive measurements:

mean 0.23, IQR [0.08, 0.35]

**11% variance due to individual differences,** **3% to verb, 16% to trial (this is encouraging!)**

**Pilot data roughly agrees:**

Gravity (pooling across objects, tasks): mean standard deviation 0.27. (first 19 subjects w/ at least 2 measurements)

Support: mean std 0.18 (first 18 w/ at least 2 measurements, only clear distinctions where one option is actually expected)

Salience: mean std 0.21 (first 18 w/ at least 2 measurements)

Estimating individual differences:

Retesting also done routinely in psychophysics/acuity estimation

e.g. Teller 1979

comparison to habituation, Adams, 1987

Colombo, Mitchell, Horowitz 1988

More fixations -> higher novelty preference

Shift rates reliable at 7 months but not 4

Familiarization (20 s at 4 months, 10 s at 7 months) and then two trials (familiar/novel)

Correlations among items relatively low (mean 0.2 at 4 months, 0.24 at 7 months)

Week-to-week reliability relatively high (.5 for # fixations and proportion time off stimulus, .4-.5 for novelty preference)

From 4 to 7 months, correlations around .3 (# fixations, novelty preference, time off stimulus, lateral bias)

IQ – novelty preference and short fixation rates

Accuracy in parameter estimation (AIPE):

- Very roughly: Need 24-31 measurements (depending on actual effect size) to have a 99% chance of getting a 90% CI on the standardized mean difference of full width equal to the standard deviation.<http://www3.nd.edu/~kkelley/publications/articles/Kelley_Rausch_2006.pdf>

- From simulation (AIPE\_sim.m): CI on mean:

o N = 16: full-width less than [1.1, 1.3, 1.4] sd (80%, 95%, 99%)

o N = 12: full-width less than [1.3, 1.5, 1.7] sd

o N = 10: [1.4, 1.7, 1.9]

Power: for 80% power, alpha=0.05, sd = 0.2 (two-sided) http://powerandsamplesize.com/Calculators/Compare-2-Means/2-Sample-Equality

- n = 39 to detect 0.6 vs. 0.5

- n = 8 to detect 0.7 vs. 0.5

sd = 0.4

- n = 125 to detect 0.6 vs. 0.5

- n = 32 to detect 0.7 vs 0.5

If averaging across all 6 trials per type per session reduces SD to 0.1:

- n = 8 to detect 0.6 vs 0.5

- n = 15 lets us detect 0.57 vs 0.5

Equivalence (checking that kids ARE performing at chance):

- 35 to tell that 0.55 is within 0.1 of 0.5 if sd=0.1

- 9 to tell that 0.5 is within 0.1 of 0.5 if sd=0.1

- 35 to tell that 0.5 is within 0.05 of 0.5

Paired comparisons within children, no correction for multiple comparisons:

- 16 to tell that 0.6 > 0.5, sd = 0.1 (i.e. 1 SD difference with 16 measurements)

- 8 to tell that 0.65 > 0.5

Seems reasonable to shoot for ~12 good sessions per child and ~12 children per approximate age group.

-> 48 included kids. (Bin 4-5 mo, 6-7 mo, 8-9 mo, 10-11 mo)